Universidade Federal Rural de Pernambuco

Unidade Acadêmica de Garanhuns

Bacharelado em Ciência da Computação

Reconhecimento de Padrões II 2016.2

Atividade 3

Prof. Tiago de Carvalho

Aluno

João Antônio Chagas Nunes

1. Introdução

O objetivo dessa atividade foi verificar a eficiência do método de extração de características derivado do *PCA* que o professor Tiago Buarque propôs no artigo *“Minimum Classification Error Principal Component Analysis”*  comparado com a forma original do *PCA.* Para isso, foram feitos testes de classificação de exemplos nas bases *The Climate Model Simulation Crashes Data Set* e *Banknote Authentication Data Set*. Utilizamos a taxa de acerto média para avaliar o resultado. Todas as classificações foram feitas com o classificadores *1*-NN, Arvore de decisão, Naive Bayes e Linear Discriminant.

2. Metodologia

PAREI AQUI

Primeiramente cada base foi separada em 10 folders e cada folder separado em parte de treino e de teste. Na base *Iris*, cada folder tinha 5 exemplos de cada classe no seu conjunto de teste e o restante dos exemplos no conjunto de treino. Já na base *Wine*, foi usado apenas 4 exemplos de cada classe no conjunto de teste e o restante dos exemplos no conjunto de treino.

A separação entre treino e teste em cada folder foi feita de forma que os exemplos escolhidos para o conjunto de teste do *i*-ésimo folder, com 0 < *i* <= 10, estavam no intervalo entre ((*x\*i*) – *x*) e (*x\*i*), onde *x* é a quantidade de exemplos de cada classe reservados para o conjunto de teste. Assim, o *j*-ésimo exemplo da classe *c* pertence ao conjunto de teste se e somente *se ((x\*i) – x) < j <= (x\*i)*.

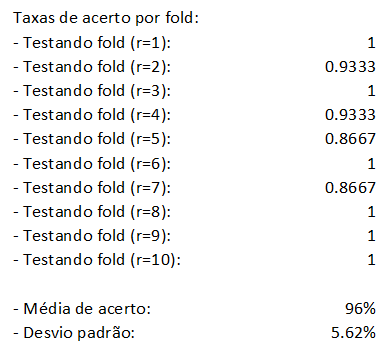
Depois que cada folder foi separado em conjunto de treino e conjunto de teste, calculamos um *PCA* para o cada folder a partir de seu conjunto de treino. Projetamos o conjunto de treino e teste de cada folder com seu devido *PCA* calculado.

Calculamos as taxas de acerto de cada folder com e sem a utilização de *PCA*, para verificar o impacto que o uso do *PCA* trás a nossos resultados. Quando utilizamos o *PCA* variamos o número de dimensões entre 1 e *d,* onde *d* é o número de dimensões da base.

No final calculamos as médias de taxa de acerto e o desvio padrão dos resultados, separando-os em sem *PCA* e com *n* dimensões do *PCA*, com *1 <= n <= d.*

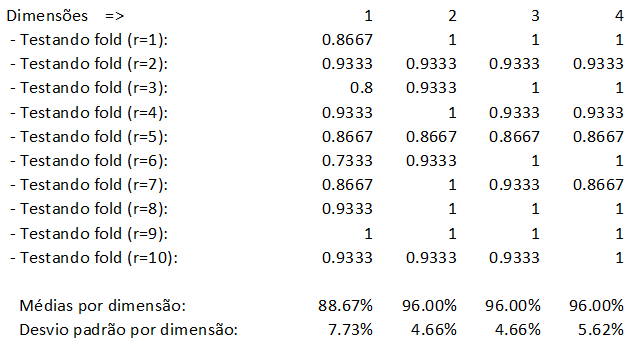
3. Resultados

O resultado do uso do *PCA* foi satisfatório, em geral conseguimos uma taxa de acerto igual ou até melhor, comparado aos mesmos teste feitos sem o uso do *PCA*, usando bem menos dimensões, o que diminui o tempo de classificação.



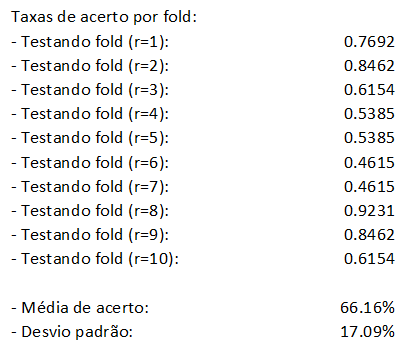
*Imagem 1* - *Resultados da base Iris, sem o uso de PCA.*

Na *Imagem 1* podemos ver a taxa de acerto que cada folder *r* teve. Os resultados obtidos na *Imagem 1* são da classificação na base Iris sem o uso de *PCA*. Podemos ver que o sem a utilização do *PCA* tivemos uma média de 96% de acerto com um desvio padrão de 5,62%.



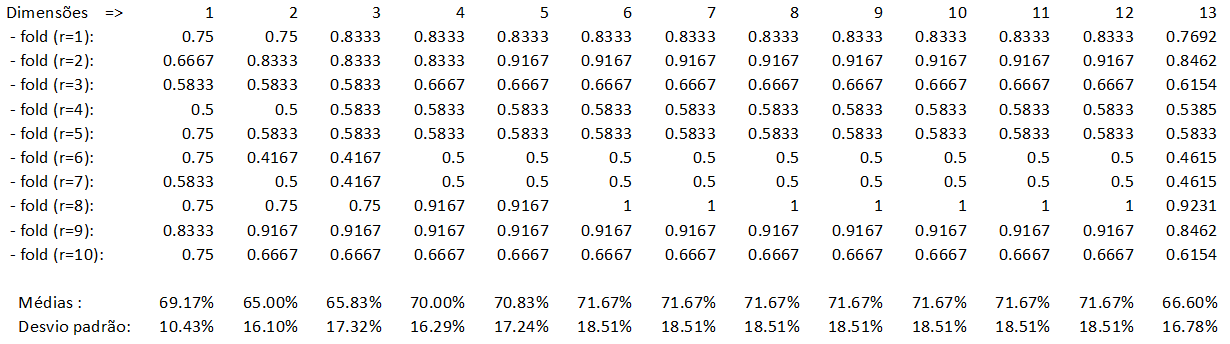
*Imagem 2 - Resultados da base Iris com a utilização do PCA.*

Na *Imagem 2* temos as taxas de acertos de cada folder da base *Iris* separados pela quantidade de dimensões usadas do *PCA*. Da pra se observar também a média de taxa de acerto e o desvio padrão de toda base separadas pelo número de dimensões usadas na classificação.



*Imagem 3 - Resultados da base Wine sem o uso do PCA.*

Na *Imagem 3* temos os resultados da base *Wine* sem o uso do *PCA*. Podemos ver que a taxa média de acerto é de 66,16% com um desvio padrão de 17,09%. Essa taxa média de acerto foi calculada usando todas as 13 dimensões da base.



*Imagem 4 - Resultados da base Wine com o uso do PCA.*

Na *Imagem 4* temos as taxas de acertos de cada folder da base *Wine* separados pelo número de dimensões usadas do *PCA*. Da pra se observar também a média de taxa de acerto e o desvio padrão de toda base separadas pelo número de dimensões usadas na classificação.

4. Conclusão

Conclui-se que a extração de características através do *PCA* pode melhorar a taxa de acerto obtida na classificação. Na base *Iris* notamos que a extração de características como *PCA* não melhorou a média de acerto da base, tivemos uma média de acerto de máxima de 96%, porém o *PCA* conseguiu obter o mesmo resultado obtido sem o seu uso, com apenas metade das dimensões da base. Ou seja, conseguimos reduzir a dimensão da base *Iris* pela metade sem prejuízos na média de acerto. Além disso, a redução da dimensão da base *Iris* também nos trouxe um desvio padrão reduzido.

Por outro lado, na base *Wine* o *PCA* conseguiu melhorar a taxa de acerto da base utilizando bem menos dimensões. A taxa de acerto da base *Wine*, que tem 13 dimensões, sem o uso do *PCA* é de 66,16%, porém usando apenas 1 dimensão do *PCA* conseguimos aumentar essa taxa de acerto em certa de 3,03% o que nos reflete em uma taxa de acerto de 69,19%. Notamos também que utilizando menos da metade de dimensões (6 dimensões) obtivemos a nossa taxa máxima de acerto.

Adicionalmente, vimos que a taxa de acerto obtida não é diretamente proporcional ao número de dimensões que usamos do *PCA*. Geralmente a taxa de acerto vária, não necessariamente para melhor, quando aumentamos o número de dimensões usadas. Assim, percebemos que não se tem uma formula pronta ditando quantas dimensões do *PCA* você deverá usar para maximizar sua taxa de acerto. O que temos que fazer é estudar nossa base e descobrir qual a melhor forma de trabalhar com ela.